ASSIST: Automated semantic assistance for translators

Serge Sharoff, Bogdan Babych
Centre for Translation Studies
University of Leeds, LS2 9JT UK
{s.sharoff,b.babych}@leeds.ac.uk

Paul Rayson, Olga Mudraya, Scott Piao
UCREL, Computing Department
Lancaster University, LA1 4WA, UK
{p.rayson,o.moudraia,s.piao}@lancs.ac.uk

Abstract

The problem we address in this paper is that of providing contextual examples of translation equivalents for words from the general lexicon using comparable corpora and semantic annotation that is uniform for the source and target languages. For a sentence, phrase or a query expression in the source language the tool detects the semantic type of the situation in question and gives examples of similar contexts from the target language corpus.

1 Introduction

It is widely acknowledged that human translators can benefit from a wide range of applications in computational linguistics, including Machine Translation (Carl and Way, 2003), Translation Memory (Planas and Furuse, 2000), etc. There have been recent research on tools detecting translation equivalents for technical vocabulary in a restricted domain, e.g. (Dagan and Church, 1997; Bennison and Bowker, 2000). The methodology in this case is based on extraction of terminology (both single and multiword units) and alignment of extracted terms using linguistic and/or statistical techniques (Déjean et al., 2002).

In this project we concentrate on words from the general lexicon instead of terminology. The rationale for this focus is related to the fact that translation of terms is (should be) stable, while general words can vary significantly in their translation. It is important to populate the terminological database with terms that are missed in dictionaries or specific to a problem domain. However, once the translation of a term in a domain has been identified, stored in a dictionary and learned by the translator, the process of translation can go on without consulting a dictionary or a corpus.

In contrast, words from the general lexicon exhibit polysemy, which is reflected differently in the target language, thus causing the dependency of their translation on corresponding context. It also happens quite frequently that such variation is not captured by dictionaries. Novice translators tend to rely on dictionaries and use direct translation equivalents whenever they are available. In the end they produce translations that look awkward and do not deliver the meaning intended by the original text.

Parallel corpora consisting of original texts aligned with their translations offer the possibility to search for examples of translations in their context. In this respect they provide a useful supplement to decontextualised translation equivalents listed in dictionaries. However, parallel corpora are not representative: millions of pages of original texts are produced daily by native speakers in major languages, such as English, while translations are produced by a small community of trained translators from a small subset of source texts. The imbalance between original texts and translations is also reflected in the size of parallel corpora, which are simply too small for variations in translation of moderately frequent words. For instance, *frustrate* occurs 631 times in 100 million words of the BNC, i.e. this gives in average about 6 uses in a typical parallel corpus of one million words.

2 System design

2.1 The research hypothesis

Our research hypothesis is that translators can be assisted by software which suggests contextual ex-

amples in the target language that are semantically and syntactically related to a selected example in the source language. To enable greater coverage we will exploit comparable rather than parallel corpora.

Our research hypothesis leads us to a number of research questions:

- Which semantic and syntactic contextual features of the selected example in the source language are important?
- How do we find similar contextual examples in the target language?
- How do we sort the suggested target language contextual examples in order to maximise their usefulness?

In order to restrict the research to what is achievable within the scope of this project, we are focussing on translation from English to Russian using a comparable corpus of British and Russian newspaper texts. Newspapers cover a large set of clearly identifiable topics that are comparable across languages and cultures. In this project, we have collected a 200-million-word corpus of four major British newspapers and a 70-million-word corpus of three major Russian newspapers for roughly the same time span (2003-2004).[1]

In our proposed method, contexts of uses of English expressions defined by keywords are compared to similar Russian expressions, using semantic classes such as persons, places and institutions. For instance, the word *agreement* in the example *the parties were frustratingly close to an agreement* = стороны были до обидного близки к достижению соглашения belongs to a semantic class that also includes *arrangement, contract, deal, treaty*. In the result, the search for collocates of близкий (close) in the context of agreement words in Russian gives a short list of modifiers, which also includes the target: до обидного близки.

2.2 Semantic taggers

In this project, we are porting the Lancaster English Semantic Tagger (EST) to the Russian language. We have reused the existing semantic field taxonomy of the Lancaster UCREL semantic analysis system (USAS), and applied it to Russian. We have also reused the existing software framework developed during the construction of a Finnish Semantic Tagger (Löfberg et al., 2005); the main adjustments and modifications required for Finnish were to cope with the Unicode character set (UTF-8) and word compounding.

USAS-EST is a software system for automatic semantic analysis of text that was designed at Lancaster University (Rayson et al., 2004). The semantic tagset used by USAS was originally loosely based on Tom McArthur's Longman Lexicon of Contemporary English (McArthur, 1981). It has a multi-tier structure with 21 major discourse fields, subdivided into 232 sub-categories.[2] In the ASSIST project, we have been working on both improving the existing EST and developing a parallel tool for Russian - Russian Semantic Tagger (RST). We have found that the USAS semantic categories were compatible with the semantic categorizations of objects and phenomena in Russian, as in the following example:[3]

poor	JJ	I1.1- A5.1- N5- E4.1- X9.1-
бедный	A	I1.1- A6.3- N5- O4.2- E4.1-

However, we needed a tool for analysing the complex morpho-syntactic structure of Russian words. Unlike English, Russian is a highly inflected language: generally, what is expressed in English through phrases or syntactic structures is expressed in Russian via morphological inflections, especially case endings and affixation. For this purpose, we adopted a Russian morpho-syntactic analyser Mystem that identifies word forms, lemmas and morphological characteristics for each word. Mystem is used as the equivalent of the CLAWS part-of-speech (POS) tagger in the USAS framework. Furthermore, we adopted the Unicode UTF-8 encoding scheme to cope with the Cyrillic alphabet. Despite these modifications, the architecture of the RST software mirrors that of the EST components in general.

The main lexical resources of the RST include a single-word lexicon and a lexicon of multi-word expressions (MWEs). We are building the Russian lexical resources by exploiting both dictionaries and corpora. We use readily available resources, e.g. lists of proper names, which are then se-

[1] Russian newspapers are significantly shorter than their British counterparts.

[2] For the full tagset, see http://www.comp.lancs.ac.uk/ucrel/usas/

[3] I1.1- = Money: lack; A5.1- = Evaluation: bad; N5- = Quantities: little; E4.1- = Unhappy; X9.1- = Ability, intelligence: poor; A6.3- = Comparing: little variety; O4.2- = Judgement of appearance: bad

mantically classified. To bootstrap the system, we have hand-tagged the 3,000 most frequent Russian words based on a large newspaper corpus. Subsequently, the lexicons will be further expanded by feeding texts from various sources into the RST and classifying words that remain unmatched. In addition, we will experiment with semi-automatic lexicon construction using an existing machine-readable English-Russian bilingual dictionary to populate the Russian lexicon by mapping words from each of the semantic fields in the English lexicon in turn. We aim at coverage of around 30,000 single lexical items and up to 9,000 MWEs, compared to the EST which currently contains 54,727 single lexical items and 18,814 MWEs.

2.3 The user interface

The interface is powered by IMS Corpus Workbench (Christ, 1994) and is designed to be used in the day-to-day workflow of novice and practising translators, so the syntax of the CWB query language has been simplified to adapt it to the needs of the target user community.

The interface implements a search model for finding translation equivalents in monolingual comparable corpora, which integrates a number of statistical and rule-based techniques for extending search space, translating words and multiword expressions into the target language and restricting the number of returned candidates in order to maximise precision and recall of relevant translation equivalents. In the proposed search model queries can be expanded by generating lists of collocations for a given word or phrase, by generating similarity classes[4] or by manual selection of words in concordances. Transfer between the source language and target language is done via lookup in a bilingual dictionary or via UCREL semantic codes, which are common for concepts in both languages. The search space is further restricted by applying knowledge-based and statistical filters (such as part-of-speech and semantic class filters, IDF filter, etc), by testing the co-occurrence of members of different similarity classes or by manually selecting the presented variants. These procedures are elementary building blocks that are used in designing different search strategies efficient for different types of translation equivalents and contexts.

The core functionality of the system is intended to be self-explanatory and to have a shallow learning curve: in many cases default search parameters work well, so it is sufficient to input a word or an expression in the source language in order to get back a useful list of translation equivalents, which can be manually checked by a translator to identify the most suitable solution for a given context. For example, the word combination *frustrated passenger* is not found in the major English-Russian dictionaries, while none of the candidate translations of *frustrated* are suitable in this context. The default search strategy for this phrase is to generate the similarity class for English words *frustrate, passenger*, produce all possible translations using a dictionary and to test co-occurrence of the resulting Russian words in target language corpora. This returns a list of 32 Russian phrases, which follow the pattern of 'annoyed / impatient / unhappy + commuter / passenger / driver'. Among other examples the list includes an appropriate translation недовольный пассажир ('unsatisfied passenger').

The following example demonstrates the system's ability to find equivalents when there is a reliable context to identify terms in the two languages. Recent political developments in Russia produced a new expression представитель президента ('representative of president'), which is as yet too novel to be listed in dictionaries. However, the system can help to identify the people that perform this duty, translate their names to English and extract the set of collocates that frequently appear around their names in British newspapers, including *Putin's personal envoy* and *Putin's regional representative*, even if no specific term has been established for this purpose in the British media.

As words cannot be translated in isolation and their potential translation equivalents also often consist of several words, the system detects not only single-word collocates, but also multiword expressions. For instance, the set of Russian collocates of бюрократия (bureaucracy) includes Брюссель (Brussels), which offers a straightforward translation into English and has such multiword collocates as *red tape*, which is a suitable contextual translation for бюрократия.

More experienced users can modify default parameters and try alternative strategies, construct

[4] Simclasses consist of words sharing collocates and are computed using Singular Value Decomposition, as used by (Rapp, 2004), e.g. *Paris* and *Strasbourg* are produced for *Brussels*, or *bus, tram* and *driver* for *passenger*.

their own search paths from available basic building blocks and store them for future use. Stored strategies comprise several elementary stages but are executed in one go, although intermediate results can also be accessed via the "history" frame. Several search paths can be tried in parallel and displayed together, so an optimal strategy for a given class of phrases can be more easily identified.

Unlike Machine Translation, the system does not translate texts. The main thrust of the system lies in its ability to find several target language examples that are relevant to the source language expression. In some cases this results in suggestions that can be directly used for translating the source example, while in other cases the system provides hints for the translator about the range of target language expressions beyond what is available in bilingual dictionaries. Even if the precision of the current version is not satisfactory for an MT system (2-3 suitable translations out of 30-50 suggested examples), human translators are able to skim through the suggested set to find what is relevant for the given translation task.

3 Conclusions

The set of tools is now under further development. This involves an extension of the English semantic tagger, development of the Russian tagger with the target lexical coverage of 90% of source texts, designing the procedure for retrieval of semantically similar situations and completing the user interface. Identification of semantically similar situations can be improved by the use of segment-matching algorithms as employed in Example-Based MT and translation memories (Planas and Furuse, 2000; Carl and Way, 2003).

There are two main applications of the proposed methodology. One concerns training translators and advanced foreign language (FL) learners to make them aware of the variety of translation equivalents beyond the set offered by the dictionary. The other application pertains to the development of tools for practising translators. Although the Russian language is not typologically close to English and uses another writing system which does not allow easy identification of cognates, Russian and English belong to the same Indo-European family and the contents of Russian and English newspapers reflect the same set of topics. Nevertheless, the application of this research need not be restricted to the English-Russian pair only. The methodology for multilingual processing of monolingual comparable corpora, first tested in this project, will provide a blueprint for the development of similar tools for other language combinations.

Acknowledgments

The project is supported by two EPSRC grants: EP/C004574 for Lancaster, EP/C005902 for Leeds.

References

Peter Bennison and Lynne Bowker. 2000. Designing a tool for exploiting bilingual comparable corpora. In *Proceedings of LREC 2000*, Athens, Greece.

Michael Carl and Andy Way, editors. 2003. *Recent advances in example-based machine translation*. Kluwer, Dordrecht.

Oliver Christ. 1994. A modular and flexible architecture for an integrated corpus query system. In *COMPLEX'94*, Budapest.

Ido Dagan and Kenneth Church. 1997. Termight: Coordinating humans and machines in bilingual terminology acquisition. *Machine Translation*, 12(1/2):89–107.

Hervé Déjean, Éric Gaussier, and Fatia Sadat. 2002. An approach based on multilingual thesauri and model combination for bilingual lexicon extraction. In *COLING 2002*.

Laura Löfberg, Scott Piao, Paul Rayson, Jukka-Pekka Juntunen, Asko Nykänen, and Krista Varantola. 2005. A semantic tagger for the Finnish language. In *Proceedings of the Corpus Linguistics 2005 conference*.

Tom McArthur. 1981. *Longman Lexicon of Contemporary English*. Longman.

Emmanuel Planas and Osamu Furuse. 2000. Multi-level similar segment matching algorithm for translation memories and example-based machine translation. In *COLING, 18th International Conference on Computational Linguistics*, pages 621–627.

Reinhard Rapp. 2004. A freely available automatically generated thesaurus of related words. In *Proceedings of LREC 2004*, pages 395–398.

Paul Rayson, Dawn Archer, Scott Piao, and Tony McEnery. 2004. The UCREL semantic analysis system. In *Proceedings of the workshop on Beyond Named Entity Recognition Semantic labelling for NLP tasks in association with LREC 2004*, pages 7–12.